ELSEVIER

Artificial Intelligence 78 (1995) 431-459

Artificial Intelligence

# Visual surveillance in a dynamic and uncertain world *

Hilary Buxton [a], Shaogang Gong [b]

[a] *School of Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton BN1 9QH, UK*

[b] *Department of Computer Science, Queen Mary and Westfield College, Mile End Road, London E1 4NS, UK*

Received September 1993; revised October 1994

**Abstract**

Advanced visual surveillance systems not only need to track moving objects but also interpret their patterns of behaviour. This means that solving the information integration problem becomes very important. We use conceptual knowledge of both the scene and the visual task to provide constraints. We also control the system using dynamic attention and selective processing. Bayesian belief networks support this and allow us to model dynamic dependencies between parameters involved in visual interpretation. We illustrate these arguments using experimental results from a traffic surveillance application. In particular, we demonstrate that using expectations of object trajectory, size and speed for the particular scene improves robustness and sensitivity in dynamic tracking and segmentation. We also demonstrate behavioral evaluation under attentional control using a combination of a static BBN TASKNET and dynamic network. The causal structure of these networks provides a framework for the design and integration of advanced vision systems.

## 1. Problem statement

Visual surveillance primarily involves the interpretation of image sequences. Advanced visual surveillance goes further and automates the detection of predefined alarm events in a given context. Whilst the definition of where and how an alarm event may occur is required, it is the intelligent dynamic scene and event discrimination which lies at the heart of advanced visual surveillance [36]. Developing a systematic methodology for the design, implementation and integration of dynamic vision systems is currently a very important research problem [3, 5, 16, 39, 51]. These methods must take into account the fact that vision is a computationally difficult problem as information available in the

* Support for the experiments reported here from ESPRIT EP2152 (VIEWS).


*SSDI* 0004-3702(95)00041-0

image does not provide a one-to-one mapping to physical objects in space. In fact, visual evidence extracted by machine-based processing is almost always subject to uncertainty and incompleteness due to noise, occlusion, and the general ill-posed nature of the inverse-perspective projection used to infer the scene from the image data. One way of overcoming some of these problems is to build in more knowledge of the scene and tasks so the primary intention of our method is to allow the representation of conceptual knowledge in a readily accessible form at all levels of visual processing. For example, in object detection and tracking in the image, we have demonstrated that it is beneficial to bring scene-based knowledge of expected object trajectories, size and speed into the interpretation process [21–23]. We have also shown that both scene and task-based knowledge allows for selective processing under attentional control for behavioral evaluation of a set of temporal events at a cluttered traffic scenario [24–26]. However, it remains to achieve all this in a tightly coupled scheme that allows for greater computational efficiency in performing multiple visual tasks.

In addition to this general requirement for integration of information in advanced visual surveillance, there are also more specific requirements for the components of the system. We have adopted the restriction from the VIEWS project that a fixed, precalibrated camera model and precomputed ground-plane geometry will be used to simplify the interpretation of the scene data in the on-line system. We also adopt the knowledge-based approach in which domain-specific models of the dynamic objects, events and behaviour are used to meet the requirement for sensitive and accurate performance. The requirements for the camera model are then to support accurate mappings from 2D to 3D and vice versa, while those for the ground-plane representation are inherited from the kind of behavioral analysis we need to support. We require both metrical and topological information to be recovered from this ground-plane representation as well as needing access to semantic and structural properties that determine the behaviour of the dynamic objects. The representation and reasoning for the dynamic objects themselves needs to support the recovery of 3D dynamic position, orientation and occupancy in the tracking as well as recognition of the object type. The requirements for the event and behaviour analysis on the other hand are less easily defined as they depend on the range of tasks to be supported. We adopt the assumption that a flexible and extensible set of behavioral evaluations and status reports may be requested which require compositional analysis in terms of the events observed. This decomposition, in which perceptual processing first recovers trajectory-based descriptions of the dynamic objects and is followed by conceptual processing, leads to a combinatorial explosion if purely data-driven control is used. Therefore, we return to the general requirement for attentional, purposive (or task-based) integration and control in advanced surveillance systems.

## 2. Introduction

The classical approach to computer vision (Marr [34]) has led to the development of sophisticated algorithms for individual visual competences. For a single visual task such as dynamic object recognition, it is possible to integrate a system using Marr's framework [38]. However, it is a clumsy approach to building systems that are required

to perform multiple tasks. In recent years, Ullman [52] has argued for the importance of integration of multiple visual routines. Of more relevance, Ballard has suggested an animate vision approach [5] for two reasons: first, vision is better understood in the context of the visual behaviours engaging the system without requiring detailed internal representations of the scene; and second, it is important to have a system framework that integrates visual processing within the task context. These arguments are supported by Bajcsy and Allen's experiments in active vision [4] and the further demonstration by Aloimonos et al. [3] that some of the intrinsically ill-conditioned processes required for the reconstruction of physical features become stable when vision processes are dynamically formulated according to the tasks. Many others have also exploited functional integration in building vision systems for a set of tasks or purposes (for example, [10,51]). In our work, we have adapted "active vision" techniques to surveillance tasks where identifying "what", "where" and "when" is just as essential for effective and efficient performance [23,26].

In visual surveillance, model-based techniques have been widely adopted for robustness and accuracy although they can be less generic. The ACRONYM system [9], for instance, used symbolic reasoning to analyse static scenes in a cycle of prediction, description, and interpretation. This strategy was effective but not sufficiently closely coupled to deliver efficient performance required for visual surveillance of dynamic scenes. Recent attempts have been made by Binford, Levitt et al. to bring about reliability and computational tractability in such systems based on applying Bayesian belief revision and decision theories [8,31]. However, they did not address the specific computational difficulties involved in the interpretation of dynamic scenes and consideration of the system purpose in the integration of visual competence. For interpretation of dynamic scenes with a wide variety of visual tasks to perform, it is very attractive to adapt dynamically the processing performed to the task at hand. Buxton and Walker [12] proposed a scheme for incorporating explicit semantic knowledge into a Query-Based Vision System for interpretation of 2D biological image sequences. However, this work did not address the question of how knowledge can be mapped onto computation in order to deliver more consistent interpretations.

In advanced visual surveillance, many different kinds of knowledge need to be represented as we are not only tracking the moving objects but also interpreting their patterns of behaviour. Pioneering work by Nagel [39] and Neumann [40] has emphasised this need to deliver conceptual or symbolic descriptions of behaviour from image sequences. Dickmanns has also developed a methodology for tightly coupled control of vision systems which builds implicit knowledge into the control loops for real-time performance to allow fixed behavioral evaluation for surveillance [17]. More immediate background here, however, is the investigation of methods for real-time knowledge-based visual surveillance systems in the ESPRIT project VIEWS. System level integration of perceptual processing with conceptual understanding of traffic scenes allowed the development of the three working demonstrators in VIEWS: airport stand area surveillance; multi-band tracker for airport ground traffic; and incident detection for road traffic scenes [14]. The conceptual processing in these systems was designed to handle missing information in the perceptual output as well as coping with behavioral variability for objects in the scenes. However, problems remain as only highly constrained feedback of information

from the conceptual processing to the perceptual level was implemented in the VIEWS demonstrators. This was mainly concentrated on occlusion handling where it was vital for the system to consistently relabel emerging vehicles [48].

In what follows, we briefly examine models and techniques for each of the computational components in a visual surveillance system. The aim of Section 3 is to identify the key competences required for an integrated system and to consider techniques that are able to provide partial or complete solutions in the light of requirements for advanced surveillance. For system integration using probabilistic models, Geman and Geman's seminal work [20] on image restoration based on stochastic relaxation initiated wide interest in addressing uncertainty. Binford et al. used Bayesian belief nets [8] and the extended influence diagram [31] to enable an "ACRONYM style" of model-based object recognition to be computationally tractable. Other proposals for applying belief theories to vision system integration include the use of Dempster–Shafer theory [7]. However, the computational complexity of such a framework suggests that it is only appropriate for conceptual evaluation where massive computation is not required. In Section 4 then, we present the theoretical basis of the Bayesian nets and corresponding belief revision mechanisms which we use in our implementations. Furthermore, Nicholson and Brady recently [41] used dynamic belief networks in addressing the "data-association" problem for vehicle monitoring. Rimey and Brown [44] applied Bayesian nets in modelling geometric constraints for active control of camera movements whilst Murino et al. [37] exploited Bayesian nets for coherent setting of multiple parameters in active control of camera operations. In Section 5, we present our approach to the similar problem of "associating" vehicles as consistent entities over time, evaluating their behaviours, and considering how to integrate the required visual competences in the context of visual surveillance. More specifically, we present some results from our experiments in motion segmentation and tracking, and attentional control in behavioral evaluation. In Section 6, we conclude our work with a summary and suggestions for future development.

## 3. Representations and computations in visual surveillance

In the VIEWS project most of the components required for an effective visual surveillance system have been identified. To simplify the run-time system we assume a precalibrated camera model, precomputed ground-plane map, as well as a set of object, event and behaviour models. We can then characterise the perceptual processing in such a system as providing track descriptions for the moving objects in the scene together with suggested object type using 2D image motion and 3D model-based vision techniques which are based on the camera, ground-plane, and object models. The conceptual processing can be characterised as taking these descriptions and then converting them into a consistent behavioral description using AI techniques based on the event and behaviour models as well as the camera and ground-plane models which define our field of view. In the following, we briefly examine five essential aspects of data representation and computation in our visual surveillance systems. These are: (1) camera models and their calibration where we emphasise the design choice for off-line fixed camera surveillance; (2) ground-plane representation where we discuss the detailed requirements for support-

ing behavioral analysis using a cellular decomposition of space; (3) object recognition where we discuss the advantages of volumetric model representations for surveillance applications; (4) tracking objects for dynamic scene interpretation where we discuss the need to fully integrate the motion analysis in model-based tracking schemes; and (5) behavioral representation and analysis where we again emphasise the need for appropriate techniques for on-line analysis and introduce cellular decomposition of time.

### *3.1. Camera models and calibration*

The design choices for building systems are fundamentally guided by the requirements of the surveillance tasks we have to accomplish. In visual surveillance, two information transformations are essential: (1) infer 3D measurements from 2D image features through inverse-perspective projection and (2) predict the existence of 2D features for 3D object hypotheses. Such transformations are determined by calibrated camera parameters. However, decisions about the type of camera calibration must take into account the fact that we use a wide-angle lens to capture the activity over a wide-area scene. Camera calibration techniques have been established across a range of requirements for accuracy and efficiency [49, 53]. However, the nature of a surveillance application based on a wide-angle, static camera means that overcoming nonlinear distortion is significant whilst dynamic calibration is less important. An existing linear model [53], which does not take into account distortion, can serve as a starting platform for establishing a more accurate system based on the radial alignment model [49]. Since these operations are computed off-line, and it is only the resulting geometry that is used on-line, the modelling can be quite elaborate to allow accurate mappings from 2D to 3D and vice versa.

### *3.2. Ground-plane representation*

The representation of the ground-plane knowledge in our system needs to be very closely bound to the behavioral analysis we need to support for our surveillance system. Although there are a wide range of spatial representation and reasoning methods reported in the literature that have been developed to support a variety of different purposes, we need a representation that is closely tailored to our requirements. In surveillance, we require both metrical information such as angles and distances between spatial primitives and topological information such as neighbours and enclosure relationships. The behavioral analysis can be facilitated by regarding the problem as interpreting the motion patterns over time within a framework provided by a static environment. It also helps to have semantic as well as structural properties made explicit in our representation as these shape the behaviour of our purposively moving objects. For example, in the road traffic domain, we need to consider not only the lane boundaries and direction of traffic flow but also the "give-way" regions.

The spatial representation and reasoning in our surveillance system, then, must support: (1) the description of the static environment in the field of view, (2) the spatial occupancy of a moving object relative to this environment including its instantaneous position, extent and the region it occupies, (3) the spatial organisation of these ob-

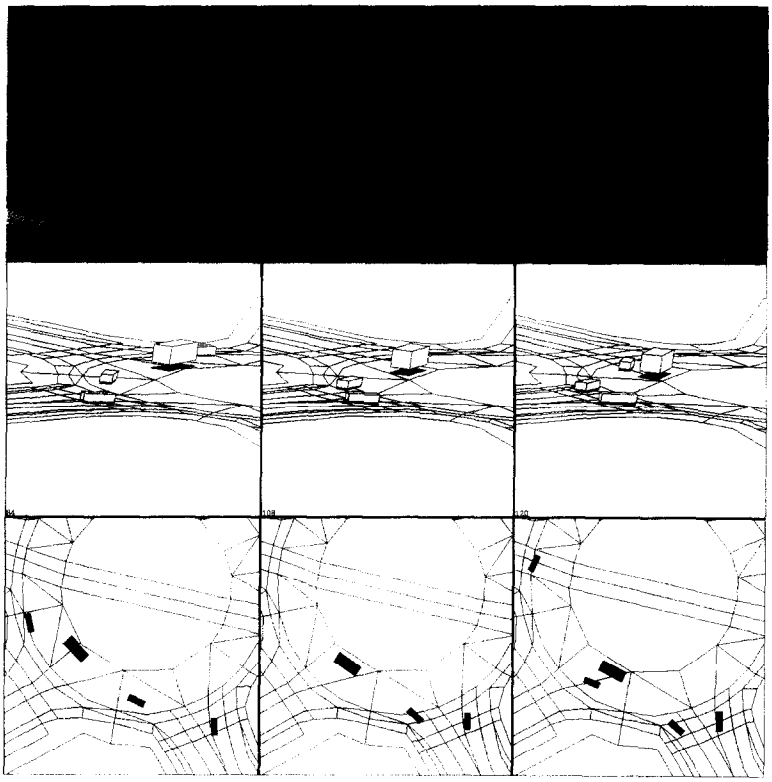

Fig. 1. The image plane, 3D space with dynamic objects, and the ground-plane representation of a traffic roundabout.

jects at a given moment with respect to the environment and each other, and (4) an understanding of what the different regions in the environment "mean" in terms of physical and semantic constraints. The semantic constraints, such as possible paths through the environment, are represented as they are effective in the interpretation of observed behaviour in line with our purposive design strategy. We also need to consider that interpretations can operate either in the image-plane or on the ground-plane projection which provides an overhead view (Fig. 1). Both types of reasoning are essentially 2D and involve both metrical and topological relationships. However, reasoning on the ground-plane will require run-time 3D model-based reasoning together with the camera model to get the necessary ground-plane projections of object position, orientation and extent. We can then integrate knowledge of the ground-plane in the motion tracking using a precomputed projection into the image-plane to provide prior expectations of object trajectory, speed and size.

Representations of space used in intermediate visual processing are concerned with supporting the immediate requirements of the task and include geometric and topological approaches. In surveillance we are concerned with perception of the moving object in the ground-plane context and need a wide-ranging model of space. An im-

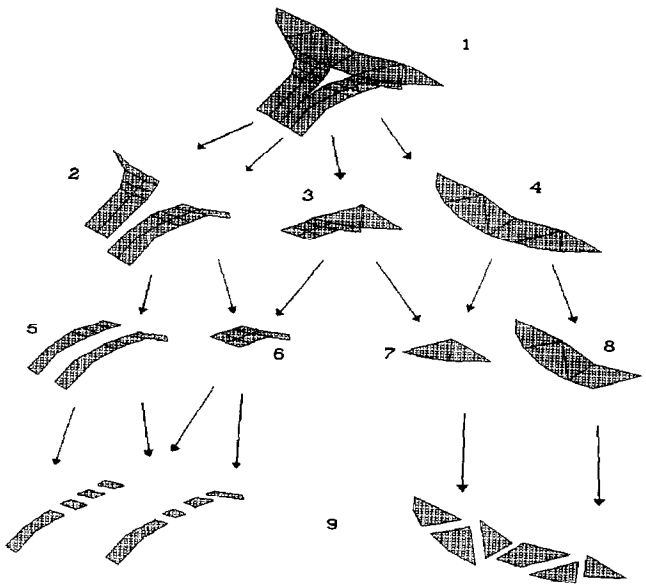


Fig. 2. Hierarchical decomposition into regions and cells representing a part of the roundabout ground-plane.

portant development has been Fleck's cellular topology [18, 19] which can support the representation of digitised spaces for edge detection and stereo matching as well as applications in qualitative physics and spatial descriptions in natural language. This representation has been extended and tailored to the needs of surveillance by Howarth and Buxton [25]. The cellular decomposition underlying this approach supports both the metrical and topological properties we require in the ground-plane framework for interpreting the behaviour of moving objects. The cells can be made to conform to the ground-plane layout and grouped into a hierarchy of regions supporting the meaningful representation of spatial context for the on-line processing of behavioral descriptions (Fig. 2). The decomposition is obtained by using a map editor to intersect (1) the road surface into (2) the entry and exit roads, (3) the turn-right zone, (4) the roundabout, (5) the lanes of the entry road, (6) the give-way zone, (7) the turning-zone, (8) the give-way-to zone, and (9) the leaf regions. The regular cells subdivide these regions to give a metrical position and support the intersection and extrusion processes used for ground-plane predictions. Spatially invariant behavioral information such as static give-way regions can then be used in the contextual indexing when we need to describe what the objects are doing in the scene. This extended cellular representation has been used to support the full event and behaviour evaluation in both passive and attentionally controlled surveillance systems [25, 26].

### *3.3. Object recognition*

The overall purpose of a visual surveillance system, as we have emphasised, is to provide meaningful descriptions of purposively moving objects. This means coherent,

effective and sufficient interpretations of dynamic behavioral patterns of 3D objects in a known scene. Therefore, the most relevant information required from a 3D object recognition and tracking system is the 3D dynamic positions, orientations and occupancies, i.e. volumes, of all the moving objects and objects that can move. Detailed information about the surface shape of individual objects are not of great concern or relevance here. This has important implications in determining the type of object representation and corresponding image descriptions.

Object recognition is one of the key visual tasks in the interpretation of dynamic activities captured in image sequences. One of the most widely used volumetric model representations is the generalised cylinder [9]. The advantages of volumetric object models are that their global properties (volumes, positions and orientations) are directly represented and easy to obtain and only a small set of values are needed to parameterise them. In contrast, surface-based representations are based on piecewise reconstruction of object surfaces, planar or curved. The surface reconstruction approach concentrates on recovering detailed information about the local geometric shape of an object. This is important in recognising objects that are primarily distinguished by their differences in shape and essential for object manipulations such as those required in robotics. However, this approach is computationally expensive and it is difficult to access global properties of objects since they are not represented and need to be inferred from the local shape information. A volumetric representation scheme, then, seems to be most appropriate for behavioral evaluation in surveillance because conceptual descriptions will need to be recovered on-line from the global properties of the objects over time.

In principle, the choice of model representation will determine the type of symbolic image description, i.e. specific extracted image features, so that they can be used to match effectively with projected geometric features on object models. The essential nature of the matching process is that the mapping between positions of the image features and the position and orientation of the models is given by a set of nonlinear functions. Although there are many model-matching techniques, it appears to us that the techniques developed so far have not adequately resolved the issue of consistent model matching in cluttered and fast changing scenes, especially the problem of invoking the correct model. We propose to avoid this problem by starting the evolution of a dynamic interpretation using the kind of simple generic volume model described above with the parameters for spatial extent, position on the ground-plane and motion refined over time by the visual evidence. A promising scheme of this kind using the "point distribution model" has been tested for tracking human figures in image sequences [6].

### *3.4. Tracking objects for dynamic scene interpretation*

For reasons of accuracy and robustness, model-based object recognition has been adopted as one of the key components in surveillance systems [17,28,56]. However, it is recognised that the temporal correlations between objects over time have not been fully incorporated into the recognition process. In other words, although model-based object tracking is generally required in the understanding of a dynamic scene with moving objects, most of the proposed schemes only address the problems of matching static 2D image descriptions to 3D object models over time. Lowe [33] and Worrall

[56] have used direct matching of image descriptions to projected descriptions of object shape, position and orientation in every frame. Marslin et al. [35] and Koller et al. [28] advance the approach by applying an independent closed form motion model for each object and match the detected static image descriptions with the motion model in each frame in order to optimise the predicted motion parameters. Schick and Dickmanns [45] further investigated a combined general shape and motion model.

However, the essence of surveillance is being able to detect and interpret change in the scene. Model-based object tracking schemes that ignore the available information about temporal changes in the image are likely to complicate the interpretation tasks in the later stages of the processing. Measuring image motion not only leads to a more compact representation of an image sequence over time, it also distinguishes noise and possible objects of interest in the scene. In general, it is well understood that the dense image motion (optic flow field) contains valuable information not only about 3D shapes of the scene but also motions in the scene. For surveillance in particular, image motion alone can be sufficient for providing information for statistical interpretations such as measuring traffic density on the roads or passenger population on train platforms. It also provides one of the most obvious bootstrapping cues for initialising object model matching when necessary.

The measurement of image motion is the primary source of detecting movement in the early stages of the interpretation process and has to be computed for effective interpretation of dynamic scenes. There have been many approaches to interpreting image motion, for example Buxton and Murray [11], Longuet-Higgins and Prazdny [32], but there has been little attempt to map dynamic image descriptions such as the optic flow field to any global 3D descriptions of objects in a model-based object recognition scheme. Only the highly computationally expensive option of detailed reconstruction of 3D surface and edges and subsequent detailed geometrical matching has been explored [38].

Most model-based object tracking techniques have adopted a simplified motion model based on the Extended Kalman Filter (EKF) in updating an object's 3D motion parameters. However, the difficulties in getting a good initial estimation of motion, which is essential to allow meaningful predictions, have again been mostly overlooked. A new probabilistic relaxation framework that reflects some of the Bayesian belief revision principles has been proposed by Kittler [27] and illustrated for matching relational structures with minimal iterations. This works by mapping evidence to expectations between different representation domains such that the most likely interpretations are obtained. This reinforces our idea that in order to overcome the limitation of the closed form EKF approach in motion estimation, a distributed belief revision approach that incorporates probability evaluations with nonlinear constraint satisfaction networks is required. It would be more appropriate for this matching problem since much weaker constraints between the evidence and expectation are imposed.

### *3.5. Behavioural representation and analysis*

When we described spatial representation and reasoning we set up the framework for the interpretation of behaviour. It is also necessary to consider how to represent and

reason about dynamic properties of moving objects as captured by the intermediate 2D and 3D tracking processes above. In general, we should note that, while there is a large area of AI and logic devoted to temporal reasoning, here we need a simple notion of time and events so we can compute the behavioral descriptions under some kind of real-time constraint. Some researchers, for example [10], have even suggested that we can dispense with internal representations and allow the world to be its own model in a "situated" action approach to modelling intelligent activity. However, we believe that we can only go part way towards this approach and do indeed need to represent the properties that are relevant for surveillance tasks. These properties will enable us to identify when a change occurs in a meaningful context, that is, here the representation of our ground-plane (scene) knowledge and other current purposively moving objects. Moving objects have trajectories in 3D space and time, 4D in general, but only 2D space and time when movement is restricted to the ground-plane. These trajectories need to be treated symbolically in order to reason about multiple object behaviour and interactions using compositional models. Differential equations [17] could be used under strict real-time requirements for surveillance but this approach makes it difficult to provide flexible and extensible conceptual descriptions of behaviour. Alternatively, we could use hidden Markov models which better capture the probabilistic nature of the visual evidence when we have a fixed, predetermined set of behaviours [21,22]. More flexible alternatives have a linguistic base with some kind of grammar based on primitive, more or less instantaneous, events and "verbs" which can be more extended relationships between objects and their environment or each other [15,29,39,40].

Surveillance is primarily about monitoring change in a local time frame sufficient to allow behaviour such as overtaking or giving way, for example, to be accomplished. This does seem to require generic, compositional analysis in terms of the state-changes or events that underlie our common sense notions of the computed conceptual descriptions. However, we propose the behavioral evaluation should be integrated into our Bayesian belief networks to reflect the probabilistic nature of evolving interpretations and provide attentional focus. Early work by Tsotsos [50] used semantic networks in the ALVEN system to integrate bottom-up, top-down and model-based processing although this cannot easily support real-time updating for dynamic vision. Nagel [39] reviewed the few projects which have tackled the problem of delivering conceptual descriptions in the road traffic domain. These include NAOS [40] and CITYTOUR [43] which allow question-answering as an off-line query process. These descriptions require composition by grammar and the underlying visual processing is also compositional in terms of a set of spatial primitives. Nagel considers the problem of on-line generation of such descriptions as we require in advanced visual surveillance. His approach goes some way towards the goal of specifying simpler (but not full natural language) conceptual descriptions in terms of motion verbs that could be effectively computed. However, we think a more situated approach has advantages for real-time performance in visual surveillance using a perceiver-centred or "deictic" frame of reference. For example, spatial deixis uses "here" and "there" and temporal deixis uses "now". Deictic reference [43] depends on knowledge of the context but can decompose and simplify the reasoning that needs to be done compared to the global "state-based" approach of traditional AI. A formalism has been developed to support reasoning from a local viewpoint and used, when required,

to construct the global behavioral descriptions which are essential in visual surveillance [26]. Here we need to understand the behaviour of individual "agents" each with their own reference frame with respect to the arbitrary global viewpoint of the camera.

We introduced the idea above that we require a more local time frame than that required to support full cognitive planning or coordinating actions as we are only observing the activity of the moving objects here. Formal approaches using well-defined languages with clear meaning for time, events, and causality, for example [2,46], as well as implementations of relational approaches to building time-maps and histories of object interactions are useful for validating and prototyping new approaches to behavioral analysis. They are not usually directly suitable for fast visual processing although Lansky's GEM (Group-Element-Model) formalism [30] can be supported by network implementations for real-time incident detection in visual surveillance as demonstrated in VIEWS. However, if we need to mix both qualitative and quantitative descriptions of time for a wider set of visual tasks, we will need new formalisms such as Fleck's cellular model of time [19] which discretises real-time in much the same way as the cellular decomposition approach for space. In cellwise-time, as developed by Fleck, each cell is a state and we can classify change as either: "state-changes", which involve sharp change; "activities", where continuous change occurs; or "accomplishments", which are composites of activity and state-change. "Episodes" are seen as composed of a starting state-change, an activity, and an end state-change in this framework. This kind of representation can be called "analogical" and has been further developed in both space and time for the representation of events and behaviour and incorporated into an experimental Bayesian network task-based control framework for surveillance by Howarth and Buxton [25,26].

## 4. Bayesian belief networks

We use the Bayesian network approach which is conceptually attractive and computationally feasible for this work. Modelling and updating the dependent relationships and their probability distributions in belief nets in such a constrained environment is relatively easy both off-line or on-line. Conceptually, we are addressing the issue of modelling a "weak" (uncertain and incomplete) constrained information retrieval process that purposively collects evidence in the image to support interpretations of dynamic behaviours in the scene. The ambiguity in interpretation means that context-dependent information integration is required to obtain more coherent descriptions of the visual evidence. Recent developments in probabilistic relaxation, belief and decision theory have provided us with a sound computational base [42].

Bayesian belief networks are Directed Acyclic Graphs (DAGs) in which each node represents an uncertain quantity using variables with multi-possible values. The arcs connecting the nodes signify the direct causal influences between the linked variables with the strengths of such influences quantified by associated conditional probabilities. If we assume a variable in the network is $X_i$, and a selection of variables $\Pi_{X_i}$ are the direct causes of $X_i$, the strengths of these direct influences are quantified by assigning the variable $X_i$ a link matrix $P(x_i \mid \Pi_{X_i})$, given any combination of instantiations of

the parent set $\Pi_{X_i}$. The conjunction of all the local link matrices of variables $X_i$ in the network (for $1 \leqslant i \leqslant n$ where $n$ is the total number of the variables) specifies a complete and consistent global model which provides answers to all the probabilistic queries. Such a conjunction is given by the overall joint distribution function over the variables $X_1, \ldots, X_n$: $P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \Pi_{X_i})$, where lower case symbols stand for a particular instantiation of the corresponding variables.[1] Then, if the behaviour of a visual process is partially defined by its processing parameters, the evaluation of these parameters will be consistent with the visual task at hand such that the task is accomplished by associating the parameters with given beliefs.

In a belief network, if we quantify the degree of coherence between the expectations ($\boldsymbol{X}$) and the evidence ($\boldsymbol{e}$) by a measure of local belief[2] $BEL(x) = P(x \mid \boldsymbol{e})$, and define belief commitments as the tentative acceptance of a subset of hypotheses that together constitute a most satisfactory explanation of the evidence at hand, then, Bayesian belief revision amounts to the updating of belief commitments by distributed local message passing operations. Instead of associating a belief measure with each individual hypothesis locally, belief revision identifies a composite set of hypotheses that best explains the evidence. We call such a set the Most-Probable-Explanation (MPE). In computational terms, this means finding the most probable instantiations of all hypothetical variables given the observation. Let $\boldsymbol{W}$ stand for all the variables concerned, inclusive of those in $\boldsymbol{e}$. Any particular instantiation of variables in $\boldsymbol{W}$ that is also consistent with $\boldsymbol{e}$ will be regarded as an *extension* or *explanation* of $\boldsymbol{e}$. The problem then is to find an extension $\boldsymbol{w}^*$ that maximises the conditional probability $P(\boldsymbol{w} \mid \boldsymbol{e})$. In other words, $\boldsymbol{W} = \boldsymbol{w}^*$ is the MPE of the evidence if $P(\boldsymbol{w}^* \mid \boldsymbol{e}) = \max_{\boldsymbol{w}} P(\boldsymbol{w} \mid \boldsymbol{e})$. Here, $\boldsymbol{w}^*$ is obtained by first locally computing the belief function for each variable $X$ mentioned above, i.e.[3] $BEL^*(x) = \max_{\boldsymbol{w}'_X} P(x, \boldsymbol{w}'_X \mid \boldsymbol{e})$ where $\boldsymbol{W}'_X = \boldsymbol{W} - X$ and second, propagating local messages. The local messages are defined as: if $X$ has $n$ parents $U_1, U_2, \ldots, U_n$ and $m$ children $Y_1, Y_2, \ldots, Y_m$, then node $X$ receives messages $\pi^*_X(u_i), i = 1, \ldots, n$, from its parents and $\lambda^*_{Y_j}(x), j = 1, \ldots, m$, from its children given by

- $\pi^*_X(u_i)$ is the probability of the most probable tail-extension of the hypothetical value $U_i = u_i$ relative to the link $U_i \rightarrow X$ and is known as an *explanation*,
- $\lambda^*_{Y_j}(x)$ is the conditional probability of the most probable head-extension of the hypothetical value $X = x$ relative to the link $X \rightarrow Y_j$, known as a *forecast*.

More precisely, given the fixed local probability $P(x \mid u_1, \ldots, u_n)$ and the best value of $X$ as $x^*$, the propagation involves:

- *Updating BEL**: for

$$F(x, u_1, \ldots, u_n) = \prod_{j=1}^{m} \lambda^*_{Y_j}(x) P(x \mid u_1, \ldots, u_n) \prod_{i=1}^{n} \pi^*_X(u_i),$$

[1] In the rest of this article, variables will always be denoted by upper case symbols and specific instantiations of the variables will be denoted by lower case symbols.

[2] In this article, all the incoming evidence will be denoted by $\boldsymbol{e}$ and be regarded as a set of instantiated variables $\boldsymbol{E}$. Symbol $\alpha$ will be used to denote a normalising constant and $\beta$ will be used for an arbitrary constant.

[3] This $BEL^*(x)$ represents the probability of the most probable extension of $\boldsymbol{e}$ that is also consistent with the hypothetical assignment $X = x$.

we have

$$BEL^*(x) = \beta \max_{u_k} F(x, u_1, \ldots, u_n), \quad 1 \leqslant k \leqslant n;$$
$$x^* = \arg\max xBEL^*(x).$$

- *Parent-bound n messages to* $U_1, \ldots, U_n$:

$$\lambda_X^*(u_i) = \max_{x, u_k:\ k \neq i} \frac{F(x, u_1, \ldots, u_n)}{\pi_X^*(u_i)}, \quad i = 1, \ldots, n.$$

- *Child-bound m messages to* $Y_1, \ldots, Y_m$:

$$\pi_{Y_j}^*(x) = \beta \frac{BEL^*(x)}{\lambda Y_j^*(x)}, \quad j = 1, \ldots, m.$$

- *Boundary conditions*: Three types of nodes set up the boundary conditions:
  (1) Anticipatory nodes: uninstantiated variables with no children. For $X$, $\lambda_{Y_j}^*(x) = [1, \ldots, 1]$.
  (2) Evidence nodes: instantiated variables. For variable $X = x'$, it is regarded as $X$ being connected with a dummy child $Z$ such that

$$\lambda_Z^*(x) = \begin{cases} 1, & \text{if } X = x', \\ 0, & \text{otherwise,} \end{cases}$$

  and other real children of $X$, $Y_1, Y_2, \ldots, Y_m$, receives the same message $\pi_{Y_j}^* = \lambda_Z^*(x)$ from $X$.
  (3) Root nodes: variables with no parents. Similarly, for each root variable, a dummy parent $U$ with permanent 1 instantiation is introduced and $P(x \mid u) = P(x) = \pi^*(x)$.

It is important to understand the conceptual essence of this propagation mechanism. For each hypothetical value of a single variable $X$, there exists a best extension of the complementary variables $W_X'$. The problem of finding the best extension of $X = x$ can be decomposed into finding the best complementary extension to each of the neighbouring variables according to the conditional independence between $X$ and the rest. This information can then be used to decide the best instantiation of $X$. The very process of this decomposition resembles the principle of optimality in dynamic programming in that it is applied recursively until it reaches the network's boundary where evidence variables have predetermined values.

## 5. Methods for integrated surveillance

In this section, we use a few specific examples to illustrate how a unified Bayesian approach can be used to overcome the problem of uncertainty and incompleteness in the visual interpretation of surveillance data by bringing task-based and scene-based knowledge into the process.

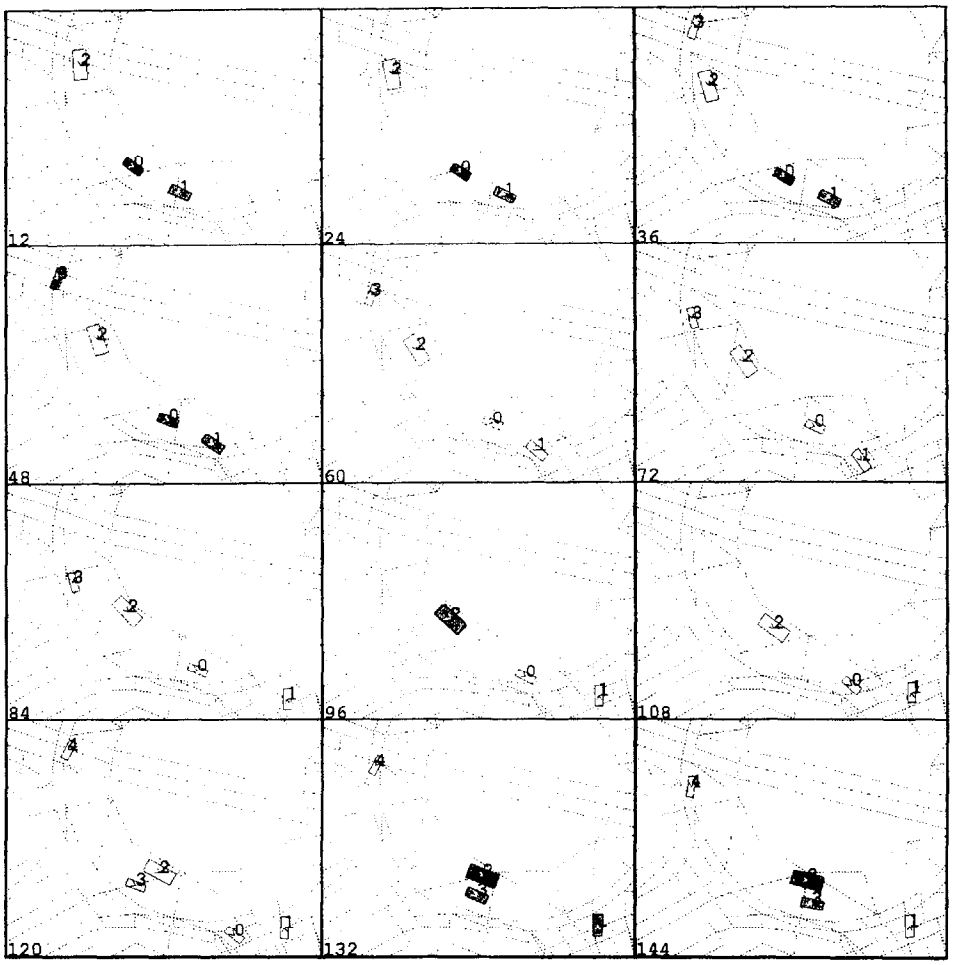


Fig. 3. Selecting which pairs of vehicles are involved in following and overtaking behaviour in an image sequence. For example, markers 0 and 1 in frame 12 are highlighted as they are close but are later ignored.

### 5.1. Behavioural evaluation

If dynamic scene interpretation in surveillance is to provide a "gestalt", the processing requires both top-down modelling of what one is looking for and bottom-up evidence for what could be appearing. The prior probabilities can be used to initialise the network and then the evidence is dynamically interpreted under the current expectations using both parent-to-child (top-down) and child-to-parent (bottom-up) updating of values in the network. We have effectively used such Bayesian networks together with a deictic representation both to create a dynamic structure to reflect the spatial organisation of the data and to measure task relatedness [26]. Howarth and Buxton integrate the behavioral evaluation and interpretation by giving a combined attentional focus for the

road traffic examplar where the behaviours of interest were "overtaking", "following", "queuing" and "unknown". For example, in Fig. 3, a simple proximity cue invokes the behavioral analysis of overtaking. A task-based Bayesian network (adapted from Rimey and Brown [44]) is used in modelling spatial and temporal relationships in order to direct the evidence collection in the image sequence.

We use a separation of preattentive (peripheral-system) and attentional (central-system) processing in our behavioral evaluation system. The set of simple low-level peripheral operators are velocity, orientation, occupied-regions, velocity-change, orientation-change, speed-difference, heading-difference, and proximity on the ground-plane. The values of these operators in effect act as cues for the more complex attentional operations such as path prediction and computing deictic spatial relationships which are used in the full evaluation. A central attentional mechanism guides the application of appropriate complex evaluation in a particular dynamic context. This attentional mechanism uses an agent-based formalism implemented by Bayesian network updating. It, first, allocates agents with unique markers to objects of interest, second, it runs the agents for the current behavioral task, and third, it collects all the results of these local agents and combines them into a global "official observer" interpretation. To understand the role of the "official observer" we can make an analogy with a sports event where the athletes are like the agents with their own individual views while a chosen spectator may have a global view. However, the "official observer" is more than this in that it both observes and directs processing so it can provide a framework for the combination of results.

The objects in our scenes have an "intrinsic front" which defines each object's frame of reference in the deictic representation. We have developed "typical-object-models" for the interpretation of behaviour using time-ordered combinations of deictic relationships between objects of interest in particular ground-plane contexts. The deictic relationships may be simply "behind"–"before", "behind and beside"–"now", and "in front"–"next". The simple peripheral operators are applied to all our segmented, tracked objects and the typical-object-model determines the specific attentional operations to be performed. It also determines which values should be saved and the set of operations to be performed on the next clock tick. The results are fed back to the appropriate agent to give task-related features for future selection. The approach here is related to that of Agre and Chapman [1] but extended to deal with many local deictic viewpoints. In this way an agent need not describe every object in the domain but only those relevant to its particular task.

*5.1.1. Experimental setup*

To combine the information that develops over time we use a dynamic form of Bayesian network (DBN) which captures the changing relationships between scene objects. The DBN is composed of temporally separated subgraphs that are interconnected using reconfigurable links (see Fig. 4). The DBN cannot model continuous time because of the graph extension process is necessarily discrete with the structure changing to reflect the temporally evolving behavioral features. This Network Expansion and Inference (NEI) algorithm consists of three steps: (1) run update-graph-structure, (2) update values of root nodes and supply evidence node values, and (3) run inference algorithm to update beliefs. If the graph structure is constant over time, it is better to

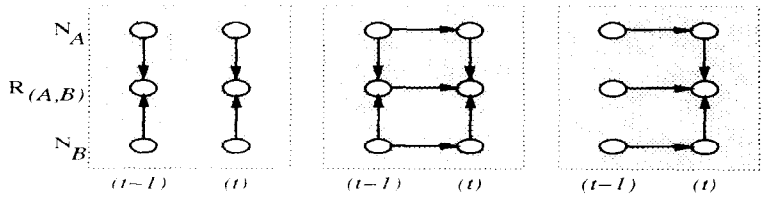


Fig. 4. Proximity relationships between vehicles $A$ and $B$ represented in a graph linking structure using analogical spatial reasoning. $N_A$ represents a vehicle node and $N_B$ a neighbouring vehicle node linked by the coefficient matrix $R_{(A,B)}$ which relates the attention given for particular event operators.

use the standard static algorithm [42] as the DBN update-graph-structure involves a more complex process of node building and node linking to return a singly connected tree. The inputs to the DBN are the Basic-Aspects which are the results delivered by the simple low-level peripheral operators that are relevant to the tasks at hand. The relevance is determined by a matrix of conditional values (i.e. ignore, watch) linking the behaviours to the operators to be used in a static BBN called the TASKNET. For example, the overtaking and following behavioral tasks have a watch value associated with the proximity operator.

The TASKNET guides the composition of the local deictic viewpoints of the selected agents to provide the official observer interpretation of the relationship. It uses a selective mechanism to ensure fast updates in the static tree structure. The input nodes represent key features relevant to the task for which it has been constructed. The output node represents the overall belief, based on the evidence collected so far, for the set of candidates consisting of the selected behavioral task, related tasks, and the default unknown.

Allocating an attentional process involves three stages: (1) the focus of attention stage ensures the TASKNET is initialized and collects evidence relevant to the selected pair of objects as well as ensuring that an agent is running on both objects; (2) the selective attention stage updates the root nodes and runs the inference algorithm; and (3) the terminate attention stage recognises when an uninteresting situation occurs and propagates a high ignore value through time which causes the allocated agents and TASKNET to terminate. Using this method, once a relationship is identified, it can then be ignored. We can only attend to a limited number of objects so a measure of the "interestingness" and the cost of performing the attentional processing is combined to provide an utility value. This value determines which object will be selected to "be watched" by the attentional processing.

The DBN and TASKNET mechanisms take one system cycle to perform and are temporally overlapped to take advantage of the current low-level peripheral operator outputs. The connection rules in the DBN (details in [24]) are the main mechanism for defining selective behaviours while the TASKNET can fine-tune the attentional mechanism to differentiate between similar surveillance tasks. The overall effect is the formation of behavioral descriptions that evolve over time to capture what is happening in the scene rather than being dependent on observing the whole episode before anything can be said to have happened.

### 5.1.2. Experimental results

As an illustration of the approach, the results for the overtaking example in Fig. 3 begin at frame 12 with two vehicles (0, 1) selected as being near each other and another vehicle (2) ignored as not interesting. At frame 24, these two vehicles are identified as involved in "following" and subsequently ignored since the behavioral task is "overtaking". In frames 36–72, two vehicles (2, 3) are selected but it is not clear from the evidence if "overtaking" or just the related "following" is taking place. It is only at frame 84 that possible "overtaking" is identified as the preferred interpretation for these two vehicles. The confirmation of the behaviour is delayed by occlusion (3 is not seen) in frames 96–108 and the re-emerging vehicle (3) is at first ignored in frame 120. The overtaking is finally confirmed to an acceptable degree of confidence at frame 132. The combination of task-related evaluation with dynamic propagation networks described above was developed in response to the requirement to deliver real-time dynamic interpretation in the behavioral analysis. Such an approach works well for relatively simple tasks like monitoring overtaking and give-way behaviour involving just two vehicles. It is also capable of multi-task evaluation and control, although it is not clear how well it would scale if the tasks involved complex multiple object interactions or plan-like behaviour.

## 5.2. Motion segmentation and tracking

In VIEWS, one of the key objectives is to segment the detected optic flow field into dynamic regions corresponding to possible moving objects and to track these regions effectively and consistently over time. Wenz [54] applied a scheme based on estimated frame displacements of the extremal loci of a bandpass filter. Similar displacement vectors are grouped into different moving regions (bounding boxes) in each frame and the similarity is defined by four parameters: (1) neighbourhood range, (2) neighbourhood displacement magnitude ratio, (3) neighbourhood orientation difference and (4) neighbourhood vector numbers. In this direct approach, these similarity parameters are set as independent constants across the entire image for computational simplicity. However, it is unable to deliver consistent interpretations in images of crowded scenes such as the traffic roundabout shown in the left picture of Fig. 5. The example frames in Figs. 9, 10 and 11 illustrate some typical defects in the sensitivity and consistency of the direct approach. We propose that scene-oriented contextual knowledge should be incorporated into the control of parameter values for more effective computation.

VIEWS uses a fixed camera for collecting visual input in each scenario. Under such static camera configurations, the three-dimensional scene layout imposes indirect, but nevertheless invariant, constraints on both possible loci of appearances, sizes, speeds of bounding boxes and the overall traffic flow. The right-hand picture of Fig. 5 illustrates the recorded motion patterns of vehicles on the roundabout over 450 frames. The *a priori* constraints on object size, speed and relationships between the parameters in the interpretation can be analysed and used to initialise the probabilities in a Bayesian net. The following correlated measures (with respect to image coordinates) are constrained probabilistically in the parameter net: (1) between object orientation and optic flow vector orientation; (2) between object size and flow vector neighbouring speed ratio, (3)

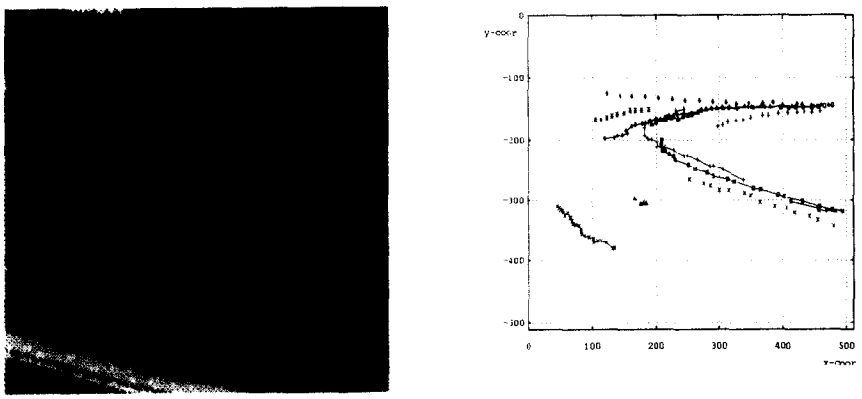

Fig. 5. Left: a traffic roundabout scenario and its traffic flow. Right: correlated spatio-temporal constraints on the movements of individual objects are imposed implicitly by this scene layout.

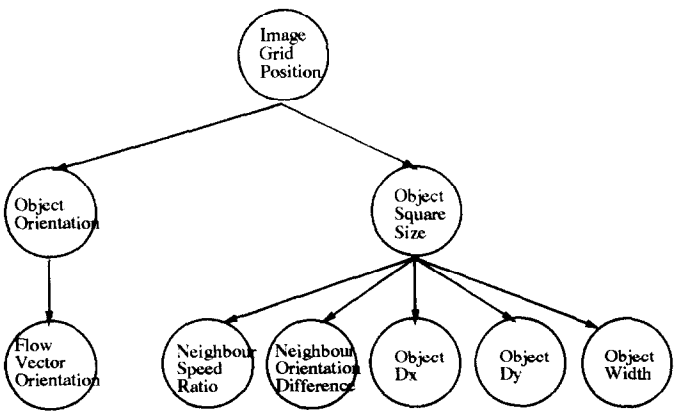


Fig. 6. A belief network that captures the dependent relationships between the scene layout and relevant measures in motion segmentation and tracking.

between neighbouring orientation difference, object d$x$, object d$y$ and object bounding box width or height. Such probabilistic constraints on the bounding boxes set up a compound network of coherent hypotheses (Fig. 6) that is modelled by a Bayesian belief network with dynamic belief revision propagation. With this approach, we regard segmenting similar flow vectors into possible moving regions in the image and tracking them down in time as providing a coherent, Most-Probable-Explanation of the detected flow fields by actively revising the distributed beliefs according to the dependent causal constraints.

### *5.2.1. Experimental setup*

The belief network in Fig. 6 has a tree structure, a special type of "singly connected" network, in order to guarantee the propagation of message passing in belief revision is tractable [42]. The image ($512 \times 512$) is divided into grids and the root node of the

tree IGP (Image Grid Position) represents the probabilistic expectation of occurrence for objects in each image grid position. Nodes OSS and OOR represent respectively the probabilistic expectations in the square size and orientation of bounding boxes in image grids. The six leaf nodes at the bottom level of the tree represent, respectively, the expectations in flow vector orientation (FVO), neighbouring vector speed ratio (NSR), orientation difference (NOD), $x$-component in object bounding box's displacement (ODX), $y$-component in bounding box displacement, and the width of a bounding box (OWD).[4] It is important to point out that first, leaf nodes are the evidence nodes and it is desirable to relate them to qualitative measures by representing relative measures between flow vectors. This is designed to overcome the instability of individual vectors in optic flow fields. Second, great effort was made to reduce the number of causal connections and the number of hypothetical variables to the minimum at the expense of approximations in the representation of certain variable nodes. This is because the computational load increases by an order of $2^n - 1$ where $n$ is the number of variable nodes in a network [13]. Third, in order to have efficient computation, it is computationally attractive to approximate any continuous variable with a set of few discrete values. Fourth, the conditional probability distribution matrices between any two nodes are usually subject to probabilistic estimation based on extensive test examples. Statistical studies in the past [13] suggest that if a well controlled number of variables are built into a Bayesian network, the estimated distribution matrices capture the general characteristics of the problem. Accurate estimation of these parameters remains one of the important factors for the computational success of a belief network. Recent studies by Spiegelhalter [47] have shown techniques for updating and learning the distribution matrices dynamically in order to provide more accuracy in their estimation.

The algorithmic steps of our approach for the segmentation and tracking of object bounding boxes from optic flow fields are: (1) Set the maximum expected number of object in a scene and initialise such a number of belief nets. (2) Set $\lambda^*_{Y_j}(x_i) = [1, \ldots, 1]$ where $X_i$ = [FVO, NSR, NOD,ODX, ODY, OWD] and $P(x \mid u) = P(x) = \pi^*(x)$ where $X$ = [IGP], then initial equilibrium of a belief tree is obtained by (a) propagating all the $\lambda$ messages upwards, (b) propagate all the $\pi$ messages downwards, (c) estimate the local beliefs throughout the tree, and (d) obtain a composite set of local instantiations of each variable that together is the best interpretation of the initial, "no evidence", condition. (3) For the first image frame, vectors are grouped according to the best value assignments associated with beliefs corresponding to their image grid position. For successive frames in the sequence, vectors are grouped according to best values, either to beliefs associated with previous tracked bounding boxes, or to beliefs associated with image grid positions. (4) For each calculated measure in the similarity test procedure, the value instantiates the associated node and revises local belief as well as other nodes' beliefs by propagation until the tree reaches equilibrium.[5] (5) Revise locally every node's best value assignment so that the bounding box will set the most

[4] The $x$, $y$ and width are measures in image coordinates. With a pre-calibrated camera and the geometry of perspective projection, they give corresponding three-dimensional measures.

[5] Since the tree is singly connected, step (4) simply means non-iterative propagation to the net boundary nodes.

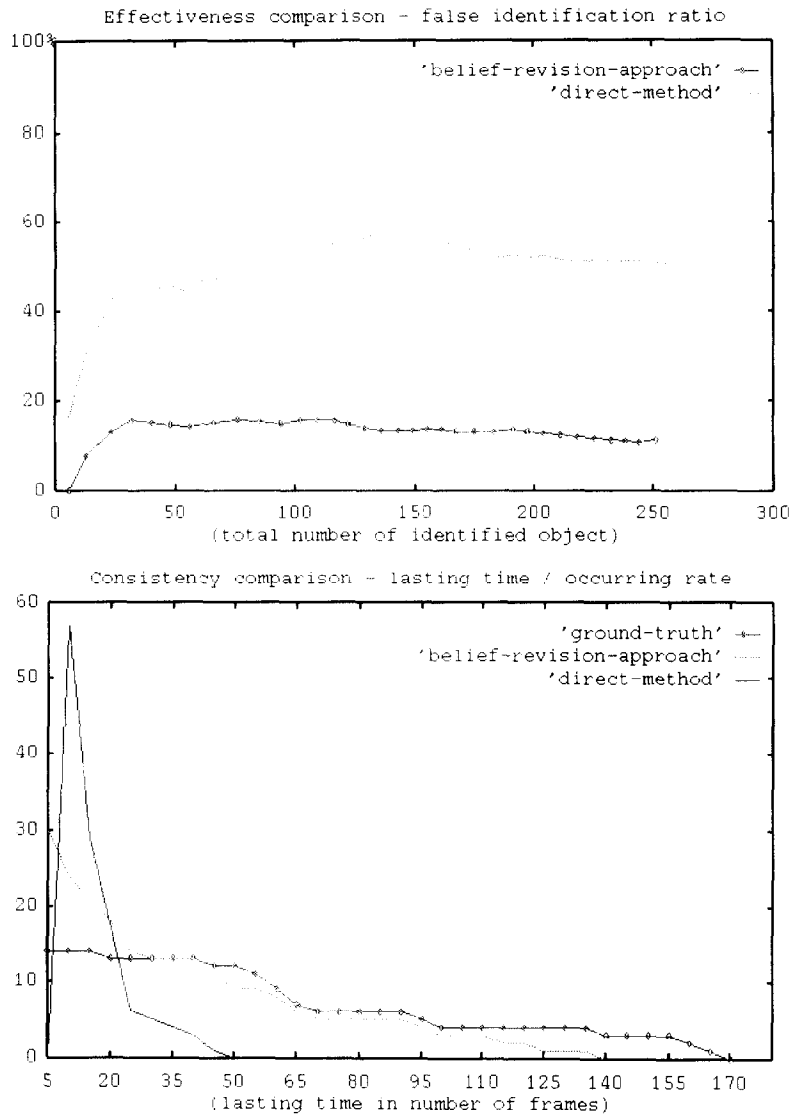


Fig. 7. Top: the false alarm rate. Bottom: the "ground truth" and detected number of objects and their duration in the scene.

probable similarity threshold values for grouping vectors that are near to its expected location in the next image frame. Repeat steps (3)–(5).

### 5.2.2. *Experimental results*

The current design of the belief network has been tested extensively on image sequences from the traffic roundabout scenario. In assessing performance, we first show the

sensitivity of the two techniques by measuring their "false alarm rate" before we measure the consistency of tracking objects over time. The false alarm rate was taken over an image sequence of 400 frames using a strict criterion of matched "true" (identified by human visual analysis on a frame-by-frame basis) and the automatically computed bounding boxes. The top graph in Fig. 7 shows the false alarm rate on both techniques over time. It gives a good indication that the belief revision approach increases true identifications significantly without introducing excessive false alarms. Throughout the whole sequence, the maximum false alarm rate from the belief revision approach is about 16%, which is below the minimum rate from the direct approach. The maximum false alarm rate of the direct approach, on the other hand, reaches 60% and its average rate is nearly 50%.

To obtain the consistency measures, we compiled the histories of tracked objects from both techniques and compare them with the "ground truth" from a 170-frame image sequence. In the bottom graph of Fig. 7, the diamond line shows the ground truth of the number of objects against their durations in the scene. For example, 1 object that has stayed for the entire 170 frames, 13 objects that have lasted for 14 frames, etc. The solid line shows that the direct approach has taken fragments of objects with long durations and tracked them as a large number of objects with very short histories. There is no object tracked for more than 50 frames, which is the basis of the poor consistency of the direct approach. In contrast, the dotted line shows that the belief revision approach provides us with more accurate measures of both the number of objects and their durations.

To estimate the computational cost, we first measure the absolute time consumption (in seconds) of both schemes over the 400-frame sequence, as seen in the two near-linear increasing lines in the top graph of Fig. 8. Although the divergence between the two lines appears to show a continuous increase of processing time in the belief revision scheme, this is due to the accumulated cost of bootstrapping belief networks over time. The frame-by-frame computational cost is more realistically given by the first-order derivative over time for those two lines, which are shown by the two step lines. This can be seen more clearly by measuring the percentage of the increased time consumption in the belief revision approach from the direct approach as seen in the bottom graph in Fig. 8. The frame-by-frame computational overhead throughout the whole sequence is below 13%, and it is worth pointing out that providing more accurate segmentation and tracking of objects instead of missing identifications will always require "extra" computational cost.

The quantitative measures presented here illustrate that: with very limited cost in computational efficiency, significant gains are obtained in effectiveness and consistency by using the belief revision technique. A more visual comparison between the two approaches can be seen in Figs. 9, 10 and 11. Three successive frames from our test sequence are shown with the results from the direct approach on the left and from the belief revision approach on the right. It is worth noticing that: first, the belief revision approach is very robust against incomplete evidence (see the tracked cyclist behind a sign post to the left-hand side of frames 145 and 150). Second, it is capable of segmenting very close moving objects (see the cyclist and the two cars close to its right). Third, in these examples, it consistently identifies all the moving objects.

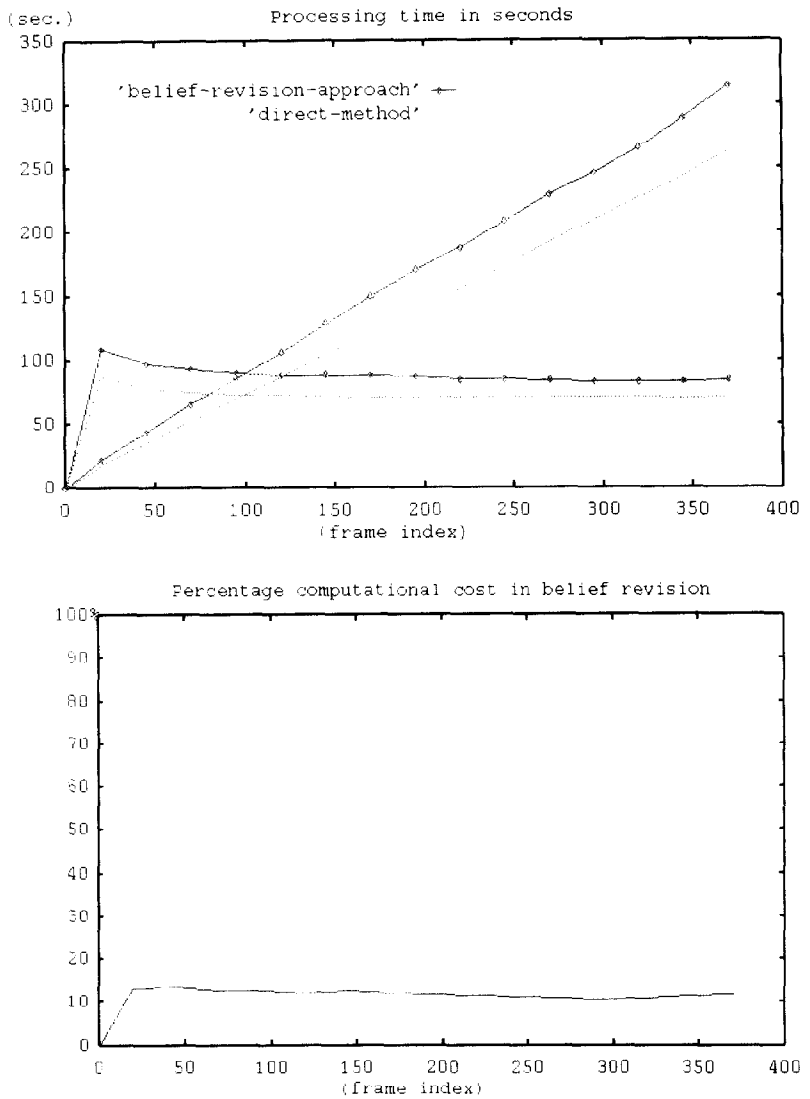


Fig. 8. Top: time in seconds for the belief revision and the direct approaches respectively, and their first-order derivatives over time. Bottom: percentage increase in the belief revision approach.

Note that these are quite cluttered traffic scenes and it is typical of these techniques to show more sensitive and robust performance on "difficult" scenes. For simple cases, direct methods can work well but have clear limitations for the full set of cases met in real-world video sequences.

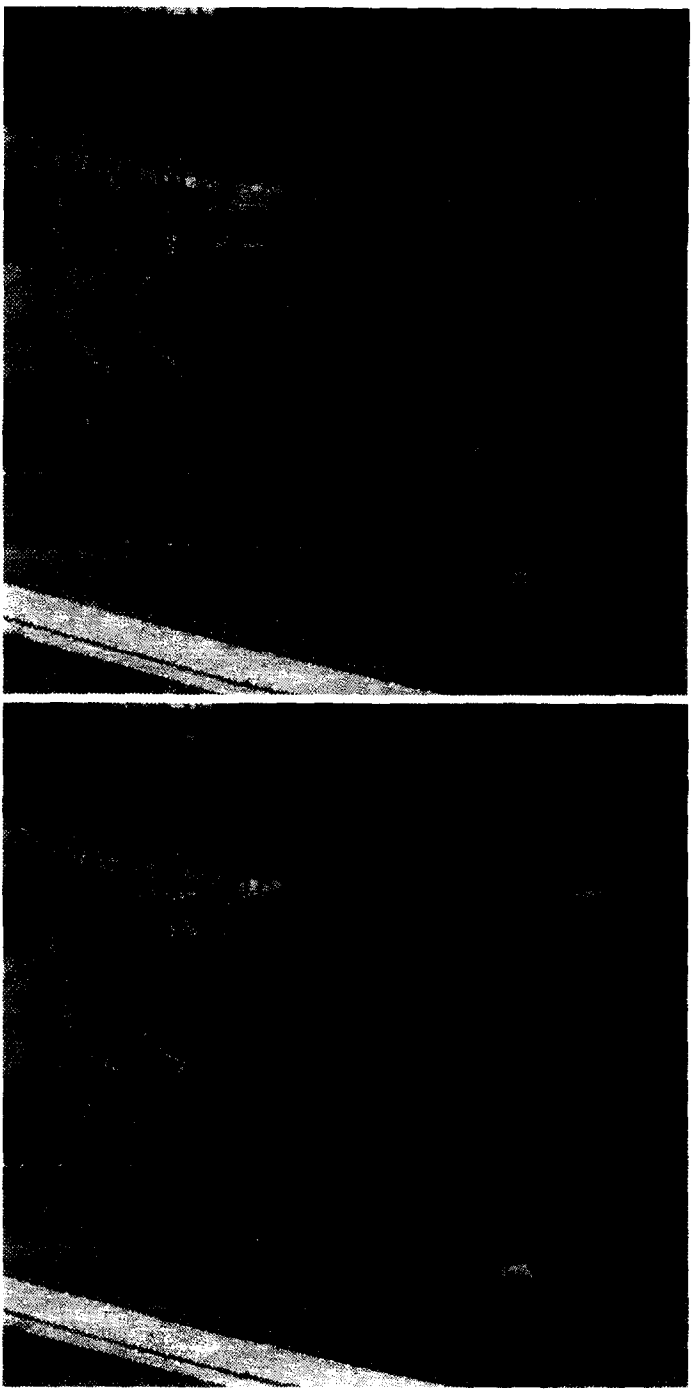

Fig. 9. Frame 140. Top: direct approach fragments object bounding boxes (compare with next two frames). Bottom: belief revision approach captures most moving objects consistently in successive frames.

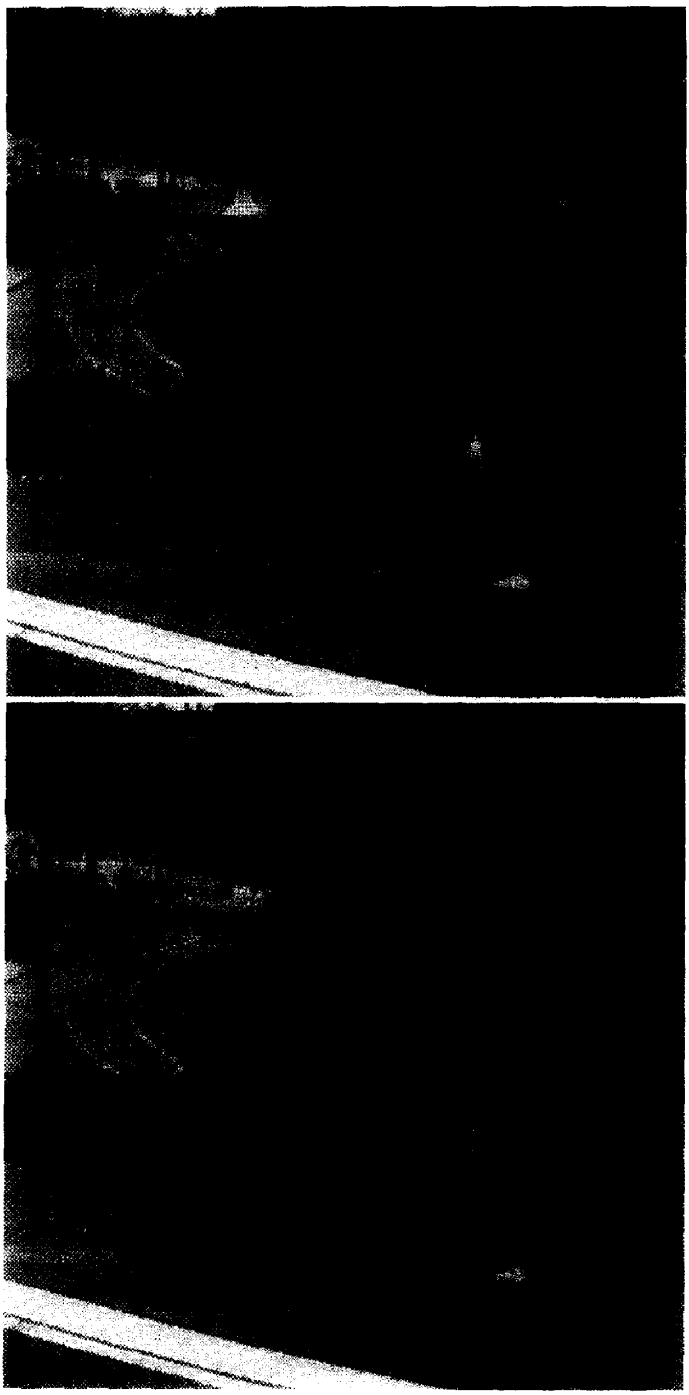

Fig. 10. Results on frame 145. Top: by the direct approach. Bottom: by the belief revision approach.

Fig. 11. Results on frame 150. Top: by the direct approach. Bottom: by the belief revision approach.

## 6. Conclusion

To summarise the arguments in the earlier sections, we are suggesting that advanced visual surveillance can benefit from taking a purposive approach to vision system design. This means using representations that are closely tailored to the surveillance tasks. We also suggest using a purposive approach in the system framework so that task-dependent processing under an active focus of attention can selectively gather evidence under the current set of expectations. In the design of a surveillance system, we point out that off-line processing, such as camera calibration or setting up of the ground-plane representation, can afford elaborate models to provide the required accuracy and functionality. However, we argue that there is also the need to find the simplest possible models of object structure and behaviour for on-line processing to accommodate the requirement for both fast interpretation and robustness with generality. For visual surveillance, we argue that the basis for the initial detection of moving objects and cues for object position and velocity should be simple visual motion measures. We also argue that the high-level interpretation, in general, requires behavioral models that are decomposable into simple primitives that can be detected in real-time and that the evolving behavioral descriptions should be computed under context-based expectations. The intermediate processing, however, poses more problems as existing model-based tracking techniques are designed for a small set of detailed models with limited dynamic updating. For more demanding surveillance tasks in which dynamic scene and event discrimination is the key [36], we propose the formulation of a new scheme within the Bayesian belief network framework as we have argued that this will provide the kind of "weak" combination of constraints appropriate for incremental shape and motion recovery in the face of uncertain and incomplete visual evidence. The experiments described in Section 5 illustrate the feasibility and computational tractability of this approach.

In Section 3, we analysed the components required in our advanced surveillance systems. We briefly reviewed progress so far and recommended the particular models and associated processing schemes that show the most promise. For example, we suggested a full radial alignment model for a fixed wide-angle surveillance camera in off-line calibration. However, if we require on-line dynamic surveillance, we would suggest simplifying the model and extending the Bayesian networks to provide a coherent model right down to the level of camera control. In the spatial and behavioral representations, we proposed cellular models as they support both qualitative and quantitative aspects of the processing as required. These models provide fast contextual indexing of computational constraints in the behavioral analysis. They have been integrated into the Bayesian belief networks to provide a framework in which interpretation evolves dynamically with a task-dependent focus of attention. In the image and object representations, we suggested using simple, reliable measures of visual motion together with volumetric models that give immediate access to the global properties of position, orientation and ground-plane motion which are required for the behavioral analysis of the moving objects. These competences, then, were all derived from a purposive strategy in visual system design.

In Section 4, we outlined the theoretical basis of Bayesian belief networks. These provide a means of performing both bottom-up, data-driven processing and top-down, expectation-driven processing in the on-line computations. Bayesian nets allow the com-

putation of the Most-Probable-Explanation of visual evidence under the expectations at all levels of abstraction in a vision system. The nodes in the parameter network are abstract entities that can be associated, for example, with simple low-level interpretations of the position and speed of bounding boxes associated with possible moving objects in the image-plane. They can equally well be associated with high-level interpretations of the kind of behaviour in which a particular object is engaged. The network updating techniques implement a fast non-recurrent solution using the current values of the nodes. The updating rules were derived from Bayesian theory which makes them very well suited to the analysis of essential visual changes in surveillance where there is always a great deal of uncertainty and incompleteness in the data. It is also important to note that the Bayesian networks allow for task-based control which is required to make the processing performed by the system selective and avoid the combinatorial explosion entailed in passive analysis. It is still important, however, to keep the networks as simple as possible and model only the essential dynamic dependencies; those that allow a rapid evaluation of the evolving spatio-temporal patterns of behaviour.

In Section 5, another aspect of the Bayesian networks becomes apparent, the ability to encode knowledge by modelling dynamic dependencies amongst the visual parameters through examples and prior probabilities of classes of interpretation. This is possible by analysis of the problem where there are obvious scene-based constraints such as traffic flow direction in certain lanes of a roundabout. It is also possible to learn these constraints and dependencies using appropriate techniques [47,55]. This can be a time consuming process but is typically computed off-line with only limited adaptive refinement on-line. The requirement to turn conceptual scene-based or task-based knowledge into a readily accessible form for real-time processing has been recognised in the past. Many hybrid schemes using both knowledge-based and numerical techniques have been proposed but would not easily support real-time systems. On the other hand, however, the kind of constraints that can be imposed using numerical techniques are rather inflexible for advanced visual surveillance where we have a lot of domain-specific knowledge. We would also argue that neural network approaches would be very difficult to develop for this class of applications. Nor do we think it possible to "evolve" solutions to such complex problems using genetic algorithms. It thus seems to us that the most promising unified framework is provided by Bayesian belief networks as we have successfully demonstrated for a set of typical advanced surveillance tasks.

## Acknowledgments

We are grateful to Richard Howarth and Andrew Smallbone for providing some of the diagrams and pictures used in this paper. We would also like to thank Bernard Buxton for his valuable comments on the initial draft of the paper.

## References

[1] P.E. Agre and D. Chapman, Pengi: an implementation of a theory of activity, in: *Proceedings AAAI-87*, Seattle, WA (1987) 268–272.

[2] J.F. Allen, Towards a general theory of action and time, *Artif. Intell.* **23** (1984) 123–154.
[3] Y. Aloimonos, I. Weiss and A. Bandopadhay, Active vision, in: *Proceedings IEEE First International Conference on Computer Vision*, London (1987) 35–54.
[4] R. Bajcsy and P. Allen, Sensing strategies, in: *Proceedings US–France Robotics Workshop* (1984).
[5] D. Ballard, Animate vision, *Artif. Intell.* **48** (1991) 57–86.
[6] A. Baumberg and D. Hogg, Learning flexible models from image sequences, Tech. Report, Research Report Series 93.36, Division of Artificial Intelligence, School of Computer Science, University of Leeds, England (1993).
[7] B. Besserer, S. Estable and B. Ulmer, Multiple knowledge sources and evidential reasoning for shape recognition, in: *Proceedings IEEE International Conference on Computer Vision*, Berlin (1993) 624–631.
[8] T.O. Binford, T.S. Levitt and W.B. Mann, Bayesian inference in model-based machine vision, in: L.N. Kanal, T.S. Levitt and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence 3*, Machine Intelligence and Pattern Recognition Series **8** (North-Holland, Amsterdam, 1989).
[9] R.A. Brooks, Symbolic reasoning among 3D models and 2D images, *Artif. Intell.* **17** (1981) 285–348.
[10] R.A. Brooks, Intelligence without reason, in: *Proceedings IJCAI-93*, Sydney, Australia (1991) 569–595.
[11] B. Buxton, D. Murray, H. Buxton and N. Williams, Structure-from-motion algorithm for computer vision on an SIMD architecture, *Comput. Phys. Commun.* **37** (1985) 273–280.
[12] H. Buxton and N. Walker, Query-based visual analysis: spatio-temporal reasoning in computer vision, *Image Vision Comput.* **6** (4) (1988) 247–254.
[13] E. Charniak, Bayesian networks without tears, *AI Mag.* **12** (4) (1991) 50–63.
[14] VIEWS Consortium, The VIEWS project and wide-area surveillance, in: *Proceedings ESPRIT Workshop at ECCV*, Genoa, Italy (1992).
[15] D.R. Corrall, A.N. Clark and A.H. Hill, Airside ground movements surveillance, in: *Proceedings NATO AGARD Symposium on Machine Intelligence in Air Traffic Management*, Berlin (1993).
[16] E.D. Dickmanns, A general dynamic vision architecture for UGV and UAV, *J. Appl. Intell.* **2** (1992) 251–270.
[17] E.D. Dickmanns, R. Behringer, F. Brudigan, F. Thomanek and V. van Holt, An all-transputer visual autobahn-autopilot/copilot, in: *Proceedings IEEE International Conference on Computer Vision*, Berlin (1993) 608–615.
[18] M. Fleck, Representing space for practical reasoning, *Image Vision Comput.* **6** (2) (1986) 75–86.
[19] M. Fleck, Boundaries and topological algorithms, Ph.D. Thesis, MIT, AI Lab., Cambridge, MA (1988).
[20] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell.* **6** (6) (1984) 721–741.
[21] S. Gong, Visual observation as reactive learning, in: *Proceedings SPIE International Conference on Adaptive and Learning Systems*, Orlando, FL (1992) 175–187.
[22] S. Gong and H. Buxton, On the expectations of moving objects, in: *Proceedings ECAI-92*, Vienna, Austria (1992) 781–785.
[23] S. Gong and H. Buxton, Bayesian nets for mapping contextual knowledge to computational constraints in motion segmentation and tracking, in: *Proceedings British Machine Vision Conference*, Guildford, England (1993) 229–238.
[24] R. Howarth, Spatial representation, reasoning and control for visual surveillance, Ph.D. Thesis, Department of Computer Science, QMW, University of London (1994).
[25] R. Howarth and H. Buxton, An analogical representation of space and time, *Image Vision Comput.* **10** (1992) 467–478.
[26] R. Howarth and H. Buxton, Selective attention in dynamic vision, in: *Proceedings IJCAI-93*, Chambery, France (1993).
[27] J. Kittler, W. Christmas and M. Petrou, Probabilistic relaxation for matching problems in computer vision, in: *Proceedings IEEE International Conference on Computer Vision*, Berlin (1993) 666–673.
[28] D. Koller, K. Daniilidis, T. Thorhallson and H.H. Nagel, Model-based object tracking in traffic scenes, in: *Proceedings European Conference on Computer Vision*, Genoa, Italy (1992) 437–452.
[29] J. Kosecka and R. Bajcsy, Cooperation of visually guided behaviours, in: *Proceedings IEEE International Conference on Computer Vision*, Berlin (1993) 502–506.

[30] A. Lansky, A representation of parallel activity based on events, structure and causality, in: *Proceedings Workshop on Reasoning about Actions and Plans* (Morgan Kaufmann, San Mateo, CA, 1986).
[31] T.S. Levitt, J.M. Agosta and T.O. Binford, Model-based influence diagrams for machine vision, in: M. Henrion, R.D. Shachter, L.N. Kanal and J.F. Lemmer, eds. *Uncertainty in Artificial Intelligence 5*, Machine Intelligence and Pattern Recognition Series **10** (North-Holland, Amsterdam, 1990).
[32] H.C. Longuet-Higgins and K. Prazdny, The interpretation of a moving retinal image, *Proc. Roy. Soc. London B* **208** (1980) 385–397.
[33] D.G. Lowe, Three-dimensional object recognition from single two-dimensional images, *Artif. Intell.* **31** (1987) 355–395.
[34] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, (Freeman, San Francisco, CA, 1982).
[35] R. Marslin, G.D. Sullivan and K.D. Baker, Kalman filters in constrained model-based tracking, in: *Proceedings British Machine Vision Conference*, Glasgow, Scotland (1991) 371–374.
[36] A. McLeod, Keeping watch on surveillance, *Image Process.* **6** (1) (1994).
[37] V. Murino, M.F. Peri and C.S. Regazzoni, Distributed belief revision for adaptive image processing regulation, in: *Proceedings European Conference on Computer Vision*, Genoa, Italy (1992) 87–91.
[38] D.W. Murray, D.A. Castelow and B.F. Buxton, From image sequences to recognised moving polyhedral objects, *Int. J. Comput. Vision* **3** (3) (1988) 107–120.
[39] H.H. Nagel, From image sequences towards conceptual descriptions, *Image Vision Comput.* **6** (1988) 59–74.
[40] B. Neumann, Natural language description of time varying scenes, in: *Semantic Structures* (Lawrence Erlbaum, Hillsdale, NJ, 1989) 167–206.
[41] A.E. Nicholson and J.M. Brady, The data association problem when monitoring robot vehicles using dynamic belief networks, in: *Proceedings ECAI-92*, Vienna, Austria (1992) 689–693.
[42] J. Pearl, *Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference* (Morgan Kaufmann, San Mateo, CA, 1988).
[43] G. Retz-Schmidt, Various views on spatial prepositions, *AI Mag.* **9** (2) (1988) 95–105.
[44] R.D. Rimey and C.M. Brown, Where to look next using a Bayes net: incorporating geometric relations, in: *Proceedings European Conference on Computer Vision*, Genoa, Italy (1992) 542–550.
[45] J. Schick and E.D. Dickmanns, Simultaneous estimation of 3D shape and motion of objects by computer vision, in: *Proceedings IEEE Workshop on Visual Motion*, Princeton, NJ (1991).
[46] Y. Shoham, *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence* (MIT Press, Cambridge, MA, 1988).
[47] D.J. Spiegelhalter and R.G. Cowell, Learning in probabilistic expert systems, in: Bayesian Statistics 4 (Oxford University Press, Oxford, 1992).
[48] A. Toal and H. Buxton, Spatio-temporal reasoning within a traffic surveillance system, in: *Proceedings European Conference on Computer Vision*, Genoa, Italy (1992) 884–892.
[49] R.Y. Tsai, A versatile camera calibration technique for high accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses, *IEEE J. Robotics Automation* **3** (4) (1987) 323–344.
[50] J.K. Tsotsos, Knowledge organisation and its role in representation and interpretation for time-varying data: the ALVEN system, *Comput. Intell.* **1** (1985) 498–514.
[51] J.K. Tsotsos, On the relative complexity of active vs. passive visual search, *Int. J. Comput. Vision* **7** (2) (1992) 127–142.
[52] S. Ullman, Visual routines, *Cognition* **18** (1984) 97–159.
[53] J. Weng, P. Cohen and M. Herniou, Camera calibration with distortion models and accuracy evaluation, *IEEE Trans. Pattern Anal. Mach. Intell.* **14** (10) (1992) 965–980.
[54] G.H. Wenz, Parallel realtime detection and tracking, M.Sc. Thesis, Department of Computer Science, QMW, University of London (1994).
[55] S.D. Whitehead and D.H. Ballard, Learning to perceive and act by trial and error, *Mach. Learn.* **7** (1991) 45–83.
[56] A.D. Worrall, R.F. Marslin, G.D. Sullivan and K.D. Baker, Model-based tracking, in: *Proceedings British Machine Vision Conference*, Glasgow, Scotland (1991) 310–318.